Composante du produit nº 82-003-X au catalogue de Statistique Canada
Rapports sur la santé

Article

Modélisation des données sur les facteurs de risque pour les données de recensement couplées : l'usage du tabac

par Claudia Sanmartin, Philippe Finès, Saeeda Khan, Paul Peters, Michael Tjepkema, Julie Bernier et Rick Burnett

Juin 2013

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

• Service de renseignements statistiques	1-800-263-1136
• Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
• Télécopieur	1-877-287-4369

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit

Le produit n° 82-003-X. au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de parcourir par « Ressource clé » > « Publications ».

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens ».

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2013

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada (http://www.statcan.gc.ca/reference/licence-fra.html).

This publication is also available in English.

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, ses entreprises, ses administrations et les autres établissements. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence ($p<0,05$)

Modélisation des données sur les facteurs de risque pour les données de recensement couplées : l'usage du tabac

par Claudia Sanmartin, Philippe Finès, Saeeda Khan, Paul Peters, Michael Tjepkema, Julie Bernier et Rick Burnett

Résumé

Contexte

Statistique Canada a entrepris une série de couplages entre les données provenant du formulaire complet du recensement de la population et des données sur les résultats pour la santé. Ces données couplées ne comportent pas d'information sur les facteurs de risque. La présente étude a pour but d'évaluer la faisabilité d'utiliser des techniques de modélisation statistique pour attribuer une situation d'usage du tabac aux participants au recensement.

Données et méthodes

On a eu recours à l'Enquête sur la santé dans les collectivités canadiennes (ESCC) de 2000-2001 pour élaborer des modèles prédictifs propres à l'âge et au sexe pouvant servir à la modélisation de la situation d'usage du tabac fondée sur les variables du Recensement de 1991. On a utilisé l'ESCC de 2002-2003 pour valider la variable modélisée. Les données tirées de l'ESCC de 2002-2003 couplées aux données provenant de la Base de données sur la morbidité hospitalière (2001-2002 à 2004-2005) ont servi à évaluer l'utilisation de la situation d'usage du tabac modélisée, comparativement à la situation d'usage du tabac autodéclarée, pour prédire les hospitalisations liées à l'usage du tabac.

Résultats

Dans le cas des modèles de personnes fumant quotidiennement, le revenu, la scolarité, l'état matrimonial, la propriété du logement et la région de naissance étaient des prédicteurs importants. Pour les modèles de personnes n'ayant jamais fumé, l'état matrimonial, la propriété du logement, l'identité autochtone et la région de naissance étaient des prédicteurs importants. La situation d'usage du tabac modélisée pour les personnes fumant quotidiennement a été associée à une cote plus élevée exprimant le risque d'une hospitalisation liée à l'usage du tabac, comparativement au fait de n'avoir jamais fumé, même après correction pour tenir compte des covariables.

Interprétation

L'étude démontre la faisabilité d'utiliser des techniques de modélisation statistique pour attribuer la situation d'usage du tabac manquante dans les données du recensement, à condition de disposer de données socioéconomiques et de données au niveau de la personne.

Mots-clés

Enquêtes sur la santé, hospitalisation, courbe ROC, facteurs socioéconomiques, modèles statistiques.

Auteurs

Claudia Sanmartin (1-613-951-6059; Claudia.Sanmartin@statcan.gc.ca), Philippe Finès (1-514-283-6847; Philippe.Finès@statcan.gc.ca), Saeeda Khan, Paul Peters, Michael Tjepkema et Julie Bernier travaillent à la Division de l'analyse de la santé de Statistique Canada, Ottawa (Ontario) K1A 0T6. Rick Burnett travaille à la Direction générale de la santé environnementale et de la sécurité des consommateurs de Santé Canada.

De plus en plus, on a recours aux données administratives pour suivre la santé de la population et mieux comprendre l'utilisation des services de santé et les résultats pour la santé. Les avantages que présentent les données administratives pour la recherche sur la santé comprennent des cohortes de taille importante fondées sur la population, de faibles coûts de collecte et un biais réduit des cas perdus de vue au suivi[1-3]. Malgré ces avantages, les données administratives renferment peu de renseignements au niveau de la personne, qui souvent se limitent aux caractéristiques démographiques, comme l'âge et le sexe, ou ne tiennent pas compte des caractéristiques socioéconomiques ou des facteurs de risque, ce qui permettrait de parvenir à une meilleure compréhension des résultats en matière de santé.

Afin de combler ces lacunes, les approches écologiques ont « annexé » aux données administratives des mesures au niveau de la région, comme les indicateurs du statut socioéconomique propres au quartier, par exemple[4-6]. Toutefois, les méthodes écologiques sont sujettes à des erreurs de classification ainsi qu'à une sous-estimation de la taille de l'effet et ne se prêtent pas aux ajustements en fonction de facteurs concurrents[7-9]. Par ailleurs, les résultats d'études fondées sur la région font ressortir non seulement les caractéristiques de la population, mais aussi celles du contexte physique et social des régions géographiques spécifiques[10].

On a également eu recours à des techniques statistiques pour effectuer des ajustements indirects relatifs aux données manquantes sur les résultats pour la santé. Par exemple, la régression « partitionnée » fait appel à des données provenant de sources auxiliaires dans les ajustements pour tenir compte des données manquantes sur les facteurs de risque[11]. Cette approche dépend de la disponibilité de données auxiliaires dans les sources de données externes ou dans les ouvrages publiés.

On a de plus en plus recours au couplage des données pour combler les lacunes dans les données administratives.

Par exemple, les données au niveau de la personne recueillies dans le cadre d'enquêtes sur la santé ont été couplées aux données tirées de dossiers hospitaliers dans des études sur les grands déterminants de l'utilisation des services hospitaliers[12-14]. Ces couplages sont riches en données sur la personne, mais la taille de l'échantillon et des facteurs associés à la couverture constituent souvent des limites en ce qui concerne les analyses portant sur des sous-groupes et des résultats moins courants.

Pour remédier à la situation, Statistique Canada a entrepris une série de projets de couplage entre les données provenant du formulaire complet du recensement et les données sur les résultats en matière de santé, notamment la mortalité, l'hospitalisation et le cancer[15,16]. Ces ensembles de données couplées offrent des données socioéconomiques exhaustives au niveau de la personne et des échantillons de taille importante, mais on y relève des lacunes en ce qui a trait aux facteurs de risque, comme l'usage du tabac et l'obésité.

La présente étude évalue la faisabilité d'utiliser des techniques de modélisation statistique pour combler les lacunes associées aux facteurs de risque, notamment l'usage du tabac, dans les données de recensement couplées[15]. En se fondant sur l'Enquête sur la santé dans les collectivités canadiennes (ESCC), on a élaboré des algorithmes prédictifs en vue de modéliser la situation d'usage du tabac à partir de variables communes à l'ESCC et au formulaire complet du Recensement de 1991. On a validé la variable résultant de ces algorithmes (usage du tabac) en comparant la contribution des situations d'usage du tabac modélisée et autodéclarée pour la prédiction d'hospitalisations liées à l'usage du tabac à partir de données d'enquêtes sur la santé et données hospitalières couplées. Cette étape était considérée comme importante, étant donné que pour évaluer l'utilité de cette approche, il est essentiel de comprendre la façon dont les données modélisées se comportent dans une analyse.

Méthodes

Source des données

On a utilisé les données de l'ESCC pour élaborer et valider des modèles prédictifs de la situation d'usage du tabac. L'ESCC est une enquête transversale qui fournit des données concernant la santé et le mode de vie de la population à domicile de 12 ans et plus des provinces et des territoires ne vivant pas en établissement, ainsi que l'utilisation faite par celle-ci des soins de santé. L'enquête exclut les membres à temps plein des Forces canadiennes et les habitants des réserves indiennes et de certaines régions éloignées. Une description détaillée de l'ESCC est disponible ailleurs[17].

L'ESCC de 2000-2001, le cycle le plus proche dans le temps de la cohorte du Recensement de 1991 couplée, a servi à élaborer les modèles prédictifs. Le taux de réponse a été de 85 %, ce qui représente un échantillon total de 131 535 personnes. L'échantillon sur lequel repose la présente étude ne tient compte que des participants à l'enquête de 25 ans et plus, soit le même critère d'âge que pour la cohorte du Recensement de 1991 couplée. Les enregistrements pour lesquels il manquait des données sur la situation d'usage du tabac ont été exclus, ce qui a donné un échantillon final de 104 204 participants à l'ESCC.

Les données tirées de l'ESCC de 2002-2003 ont servi à la validation externe des modèles prédictifs. Le taux de réponse a été de 81 %, ce qui représente un échantillon total de 134 072 personnes. Des exclusions similaires ont donné lieu à un échantillon de validation final de 107 398 personnes.

Les données provenant de l'ESCC de 2002-2003, couplées à la Base de données sur la morbidité hospitalière (BDMH) (2001-2002 à 2004-2005), ont servi à évaluer les associations entre les situations d'usage du tabac modélisée et autodéclarée et les hospitalisations liées à l'usage du tabac. La BDMH est un ensemble de données administratives au niveau de la personne portant sur les hospitalisations dans la plupart des hôpitaux de soins actifs et dans certains hôpitaux psychiatriques, de soins de longue durée et de réadaptation au Canada[18]. Le couplage des données a été effectué pour les participants à l'ESCC vivant à l'extérieur du Québec qui ont consenti au couplage et qui ont fourni un numéro d'assurance-maladie personnel valide (n=81 364). Des exclusions similaires ont été appliquées aux données couplées (25 ans et plus; données manquantes sur l'usage du tabac), ce qui a produit un échantillon final de 52 396 personnes. Des détails concernant le couplage des données figurent ailleurs[12,19].

Élaboration de modèles prédictifs

Des modèles distincts ont été élaborés pour prédire les estimations pour deux catégories d'usage du tabac : les *personnes fumant quotidiennement* et les *personnes n'ayant jamais fumé*. Les catégories d'usage du tabac ont été calculées à partir des données autodéclarées dans l'ESCC[20]. Les personnes fumant quotidiennement étaient celles qui avaient déclaré fumer sur une base quotidienne dans le cadre de l'enquête (1=oui, 0=non). Les personnes n'ayant jamais fumé étaient celles qui avaient déclaré n'avoir jamais fumé ou avoir fumé moins de 100 cigarettes au cours de leur vie (1=oui, 0=non). Les tentatives de prédiction concernant les anciens fumeurs ont été infructueuses, les modèles ne permettant pas de faire de distinction entre les personnes fumant quotidiennement, les personnes n'ayant jamais fumé et les anciens fumeurs.

Aux fins de la prédiction de la situation d'usage du tabac, les variables de l'ESCC devaient avoir été incluses dans le recensement (formulaire complet) et on devait avoir démontré ou fait l'hypothèse qu'elles étaient associées à l'usage du tabac. Dans la mesure du possible, les variables de l'ESCC ont été codées pour correspondre aux définitions des variables du recensement. Des variables économiques, sociodémographiques, de logement et ethnoculturelles ont servi à prédire la situation d'usage du tabac (tableau 1).

Des modèles de régression logistique multivariée ont été élaborés pour prédire la probabilité d'être une personne fumant quotidiennement et une personne n'ayant jamais fumé. Les analyses préliminaires ayant révélé une variabilité des facteurs associés à la situation d'usage du tabac selon le sexe et le groupe d'âge, on a élaboré des modèles propres aux combinaisons de sexe et de groupe d'âge. L'échantillon complet de l'étude a été utilisé dans les deux ensembles de modèles, afin de pouvoir attribuer à chaque participant à l'ESCC une estimation de la probabilité d'être une personne fumant quotidiennement et une estimation de la probabilité d'être une personne n'ayant jamais fumé. On a eu recours à la technique pas à pas, afin d'assurer la sélection de la liste de variables la plus courte possible pour chaque combinaison de sexe et de groupe d'âge; les variables ont été incorporées dans le modèle par ordre décroissant d'importance de la signification. Des poids d'enquête ont été utilisés, et la méthode du *bootstrap* (rééchantillonnage) a été appliquée aux modèles de régression multivariée finaux, afin de tenir compte du plan de sondage complexe de l'ESCC. Les modèles ont été élaborés au moyen de PROC LOGISTIC en SAS, version 9.1.

Des seuils propres au modèle ont été établis pour répartir les participants à l'enquête en catégories d'usage du tabac. Plus particulièrement, on a produit des courbes caractéristiques de la performance d'un test (courbes ROC pour *Receiver Operating Characteristic curves*)[21-24] pour déterminer les seuils de probabilité optimaux propres à l'âge et au sexe. Une personne était identifiée comme étant un cas positif si ses estimations de la probabilité d'être une personne fumant quotidiennement (ou une personne n'ayant jamais fumé) dépassaient les seuils optimaux. Les seuils optimaux ont été produits pour atteindre un équilibre entre les faux positifs et les faux négatifs, dans le but de réduire les premiers. Compte tenu de la taille importante des échantillons associés au recensement, on obtient un modèle plus précis en se concentrant sur

Tableau 1
Répartition en pourcentage (pondérée) de certaines caractéristiques de l'échantillon de l'étude, population à domicile de 25 ans et plus, Canada, 2000-2001

Caractéristiques choisies	Pourcentage	Intervalle de confiance de 95 %	
		de	à
Situation d'usage du tabac			
N'a jamais fumé	41,4	41,0	41,9
Fume quotidiennement	26,1	25,7	26,5
Autre	32,5	32,0	32,9
Sexe			
Femmes	51,2	51,1	51,4
Hommes	48,8	48,6	48,9
Groupe d'âge			
25 à 44 ans	46,4	46,3	46,6
45 à 64 ans	35,7	35,6	35,8
65 ans et plus	17,9	17,8	17,9
État matrimonial			
Célibataire, jamais marié(e)	13,7	13,5	14,1
Marié(e) / en union libre	71,0	70,4	71,3
Séparé(e)	3,0	2,8	3,1
Divorcé(e)	5,7	5,6	6,0
Veuf(veuve)	6,6	6,4	6,7
Niveau de scolarité			
Pas de diplôme d'études secondaires	23,4	23,0	23,8
Diplôme d'études secondaires	37,9	37,4	38,3
Études postsecondaires partielles	19,8	19,4	20,2
Diplôme d'études universitaires	19,0	18,5	19,4
Quintile de revenu			
Inférieur	17,4	16,9	17,7
Moyen-inférieur	17,9	17,5	18,2
Moyen	18,1	17,8	18,5
Moyen-supérieur	18,1	17,7	18,4
Supérieur	18,9	18,5	19,3
Données manquantes	9,7	9,5	10,0
Situation d'emploi			
Personne occupée	63,9	63,6	64,3
Personne inactive	32,6	32,3	33,0
Personne en chômage	3,4	3,3	3,6
Propriété du logement			
Non	27,5	27,1	28,1
Oui	72,5	71,9	72,9
Taille du ménage (nbre de personnes)			
Une	14,8	14,6	15,2
Deux	34,8	34,5	35,3
Trois	18,8	18,4	19,2
Quatre	19,4	19,0	19,8
Cinq	8,0	7,7	8,2
Six ou plus	4,2	3,9	4,4
Ascendance autochtone			
Non	97,2	97,1	97,4
Oui	2,8	2,6	2,9
Appartenance à une minorité visible			
Non-blanc	14,1	13,6	14,5
Blanc	85,9	85,5	86,4
Région de naissance			
Canada	76,2	75,9	76,8
Autre pays d'Amérique du Nord	1,3	1,2	1,4
Amérique du Sud / Amérique centrale / Caraïbes	2,5	2,3	2,7
Europe et Océanie	11,0	10,7	11,3
Afrique	1,2	1,1	1,4
Asie	7,7	7,4	8,0
Indicateur rural/urbain			
Région rurale (agricole et non agricole)	18,3	17,7	18,8
Petite région urbaine (moins de 30 000 habitants)	15,0	14,7	15,5
Région urbaine (de 30 000 à 99 999 habitants)	9,7	9,3	10,0
Région urbaine (de 100 000 à 499 999 habitants)	11,4	11,1	11,7
Région urbaine (500 000 habitants et plus)	45,6	45,0	46,1

Source : Enquête sur la santé dans les collectivités canadiennes, 2000-2001.

Tableau 2
Variables prédisant la situation d'usage du tabac (en ordre d'importance), selon le sexe et le groupe d'âge, population à domicile de 25 ans et plus, Canada, 2000-2001

	Hommes			Femmes		
Modèle de situation d'usage du tabac	**25 à 44 ans**	**45 à 64 ans**	**65 ans et plus**	**25 à 44 ans**	**45 à 64 ans**	**65 ans et plus**
Personne fumant quotidiennement						
Variables explicatives	Scolarité	Propriété du logement	État matrimonial	Scolarité	Région de naissance	Propriété du logement
	Propriété du logement	Scolarité	Quintile de revenu	Région de naissance	Propriété du logement	Origine ethnique
	Région de naissance	Région de naissance	Taille du ménage	Propriété du logement	Scolarité	Scolarité
	État matrimonial	État matrimonial	Région de naissance	État matrimonial	État matrimonial	État matrimonial
	Situation d'emploi	Quintile de revenu	Propriété du logement	Quintile de revenu	Quintile de revenu	Région de naissance
	Quintile de revenu	Nombre de chambres à coucher	Scolarité	Taille du ménage	Taille du ménage	Quintile de revenu
	Ascendance autochtone	Situation d'emploi		Ascendance autochtone	Nombre de chambres à coucher	Taille du ménage
	Taille du ménage	Ascendance autochtone		Situation d'emploi	Ascendance autochtone	
		Taille du ménage		Origine ethnique	Origine ethnique	
		Indicateur rural/urbain				
ASC	0,702	0,677	0,643	0,729	0,672	0,586
Seuil optimal	0,305	0,257	0,108	0,269	0,223	0,105
% de prédictions correctes (seuil optimal)	63,5 %	62,2 %	59,9 %	66,5 %	61,9 %	54,4 %
% de prédictions correctes (seuil optimal +/- 0,05)	46,0 %	42,1 %	12,5 %	50,8 %	39,9 %	12,8 %
% de prédictions correctes (seuil optimal +/- 0,10)	37,2 %	26,1 %	2,8 %	38,6 %	22,7 %	1,1 %
Personne n'ayant jamais fumé						
Variables explicatives	Scolarité	Scolarité	Région de naissance	Région de naissance	Région de naissance	Région de naissance
	Région de naissance	Région de naissance	Scolarité	Scolarité	Propriété du logement	État matrimonial
	Propriété du logement	Propriété du logement	Situation d'emploi	Propriété du logement	État matrimonial	Propriété du logement
	Situation d'emploi	Taille du ménage	Propriété du logement	Taille du ménage	Scolarité	Origine ethnique
	État matrimonial	État matrimonial	État matrimonial	État matrimonial	Nombre de chambres à coucher	Ascendance autochtone
	Nombre de chambres à coucher	Situation d'emploi	Ascendance autochtone	Ascendance autochtone	Ascendance autochtone	Quintile de revenu
	Ascendance autochtone	Nombre de chambres à coucher	Origine ethnique	Origine ethnique	Origine ethnique	Taille du ménage
	Origine ethnique	Quintile de revenu	Taille du ménage	Quintile de revenu	Quintile de revenu	Indicateur rural/urbain
	Taille du ménage	Ascendance autochtone	Indicateur rural/urbain	Situation d'emploi	Taille du ménage	Nombre de chambres à coucher
	Quintile de revenu	Indicateur rural/urbain				
		Origine ethnique				
ASC	0,679	0,659	0,596	0,703	0,637	0,591
Seuil optimal	0,574	0,698	0,731	0,552	0,549	0,385
% de prédictions correctes (seuil optimal)	62,3 %	60,9 %	57,3 %	64,6 %	59,4 %	56,3 %
% de prédictions correctes (seuil optimal +/- 0,05)	46,4 %	41,6 %	21,7 %	50,2 %	36,0 %	28,0 %
% de prédictions correctes (seuil optimal +/- 0,10)	38,3 %	27,4 %	10,3 %	37,8 %	24,7 %	12,0 %

ASC = aire sous la courbe ROC

les vrais positifs, même si un nombre important de faux négatifs sont produits.

La validation du modèle a été évaluée selon l'aire sous la courbe ROC (ASC), laquelle est le tracé de la sensibilité par rapport à 1 – spécificité (c.-à-d. un moins la spécificité). En outre, on a calculé le pourcentage de cas prédits correctement en comparant la situation d'usage du tabac autodéclarée à celle qui était obtenue dans les modèles.

Attribution de la situation d'usage du tabac

Les probabilités prédites pour les catégories d'usage du tabac de personne fumant quotidiennement et de personne n'ayant jamais fumé ont servi à classer chaque personne tant dans la catégorie des personnes fumant quotidiennement que dans celle des personnes n'ayant jamais fumé, d'après les seuils propres au sexe et à l'âge. Les deux systèmes de classification ont par la suite servi à l'attribution de la situation d'usage du tabac définitive :

Personne n'ayant jamais fumé	Personne fumant quotidiennement	
	Oui	Non
Oui	Impossible à classer	Personne n'ayant jamais fumé
Non	Personne fumant quotidiennement	Autre

Par exemple, les participants à l'enquête pour qui la probabilité d'être une personne n'ayant jamais fumé dépassait le seuil propre au groupe âge-sexe et la probabilité d'être une personne fumant quotidiennement était inférieure au seuil ont été classés dans la catégorie des personnes n'ayant jamais fumé. Les participants à l'enquête pour qui les probabilités les plaçaient à la fois dans la catégorie des personnes fumant quotidiennement et dans celle des personnes n'ayant jamais fumé ont été considérés comme *impossibles à classer* et supprimés de l'analyse. Les participants à l'enquête entrant dans la catégorie *autres* n'ont été considérés ni comme fumant quotidiennement ni comme n'ayant jamais fumé; il pouvait s'agir de fumeurs occasionnels ou d'anciens fumeurs, ou encore de faux négatifs.

Des bandes de seuils, calculées selon seuil optimal +/-0,05 ou +/-0,10, ont été produites aux fins d'analyses de sensibilité. Si les probabilités prédites étaient supérieures (inférieures) au seuil le plus élevé (le moins élevé), les participants à l'enquête étaient considérés représenter des cas positifs (négatifs) quant à la situation de personne fumant quotidien-

Tableau 3
Rapports de cotes non corrigés et corrigés reliant la situation d'usage du tabac autodéclarée et modélisée aux hospitalisations de courte durée liées à l'usage du tabac, population à domicile de 25 ans et plus, Canada, à l'exclusion du Québec, 2002-2003 à 2004-2005

	Personne fumant quotidiennement		
	Rapport de cotes	Intervalle de confiance de 95 %	
		de	à
Situation d'usage du tabac autodéclarée			
Non corrigée (situation d'usage du tabac seulement)	1,71*	1,38	2,11
Partiellement corrigée (âge et sexe)	2,41*	1,92	3,02
Entièrement corrigée†	2,19*	1,74	2,77
Situation d'usage du tabac modélisée			
Seuil optimal			
Non corrigée (situation d'usage du tabac seulement)	1,81*	1,53	2,12
Partiellement corrigée (âge et sexe)	1,97*	1,65	2,34
Entièrement corrigée†	1,30*	1,04	1,63
Seuil optimal +/- 0,05			
Non corrigée (situation d'usage du tabac seulement)	2,53*	1,87	3,43
Partiellement corrigée (âge et sexe)	2,69*	1,96	3,67
Entièrement corrigée†	1,52*	1,02	2,26
Seuil optimal +/- 0,10			
Non corrigée (situation d'usage du tabac seulement)	2,99*	1,89	4,73
Partiellement corrigée (âge et sexe)	2,72*	1,70	4,36
Entièrement corrigée†	1,07	0,61	1,85

* valeur significativement différente de celle observée pour le groupe des personnes « n'ayant jamais fumé » ($p < 0,05$)
† âge, sexe, ascendance autochtone, minorité visible, état matrimonial, niveau de scolarité, revenu, emploi
Sources : Enquête sur la santé dans les collectivités canadiennes, 2002-2003; Base de données sur la morbidité hospitalière, 2001-2002 à 2004-2005.

Ce que l'on sait déjà sur le sujet

- Les données administratives sont de plus en plus utilisées pour suivre la santé de la population et comprendre l'utilisation des services de santé et les résultats pour la santé.
- On a recours au couplage de données pour combler les lacunes dans les données administratives.
- Dans certaines données couplées, il subsiste des lacunes dans l'information (p. ex., sur les facteurs de risque).

Ce qu'apporte l'étude

- La présente étude évalue la faisabilité d'utiliser des techniques de modélisation statistique pour combler les lacunes, notamment à l'égard de la situation d'usage du tabac, dans les données de recensement (formulaire complet) couplées.
- On a élaboré des algorithmes prédictifs aux fins de la modélisation de la situation d'usage du tabac, à partir de l'Enquête sur la santé dans les collectivités canadiennes.
- Une analyse de régression a démontré la viabilité d'utiliser la variable modélisée de l'usage du tabac pour examiner les associations entre la situation d'usage du tabac et les hospitalisations liées à l'usage du tabac.

nement et de personne n'ayant jamais fumé, ce que l'on estime être fondé étant donné que la valeur de résultat prédite ne correspond pas au produit final de l'analyse, mais plutôt à l'attribution de la situation d'usage du tabac appropriée.

Application de la situation d'usage du tabac modélisée

Les données de l'ESCC de 2002-2003 et les données hospitalières de 2001-2002 à 2004-2005 couplées ont servi à déterminer comment la variable modélisée de l'usage du tabac se comporte dans les analyses des issues en santé. Les objectifs étaient les suivants : 1) comparer l'association entre la situation d'usage du tabac et les hospitalisations liées à cet usage, pour les situations d'usage modé-

Tableau 4
Rapports de cotes corrigés reliant la situation d'usage du tabac autodéclarée et modélisée et certaines caractéristiques aux hospitalisations de courte durée liées à l'usage du tabac, population à domicile de 25 ans et plus, Canada, à l'exclusion du Québec, 2002-2003 à 2004-2005

Caractéristiques choisies	Modèle A (usage du tabac autodéclaré)			Modèle B (usage du tabac modélisé - seuil optimal)			Modèle C (usage du tabac modélisé - seuil optimal +/- 0,05)			Modèle D (usage du tabac modélisé - seuil optimal +/- 0,10)		
	Rapport de cotes corrigé	Intervalle de confiance de 95 %		Rapport de cotes corrigé	Intervalle de confiance de 95 %		Rapport de cotes corrigé	Intervalle de confiance de 95 %		Rapport de cotes corrigé	Intervalle de confiance de 95 %	
		de	à		de	à		de	à		de	à
Situation d'usage du tabac												
N'a jamais fumé†	1,00	...	...	1,00	...	...	1,00	...	...	1,00	...	...
Fume quotidiennement	2,19*	1,74	2,77	1,30*	1,04	1,63	1,52*	1,02	2,26	1,07	0,61	1,85
Autre	1,51*	1,28	1,79	1,21	0,95	1,53	0,88	0,35	2,17	0,99	0,55	1,80
Sexe												
Hommes†	1,00	...	...	1,00	...	...	1,00	...	...	1,00	...	...
Femmes	0,51*	0,44	0,60	0,47*	0,40	0,54	0,47*	0,40	0,55	0,46*	0,39	0,54
Groupe d'âge												
25 à 44 ans†	1,00	...	...	1,00	...	...	1,00	...	...	1,00	...	...
45 à 64 ans	4,72*	3,18	7,02	4,77*	3,19	7,14	4,73*	3,14	7,14	4,65*	3,07	7,04
65 à 75 ans	9,46*	6,43	13,91	8,94*	5,83	13,70	8,46*	5,35	13,36	8,38*	5,19	13,53
État matrimonial												
Marié(e) / en union libre†	1,00	...	...	1,00	...	...	1,00	...	...	1,00	...	...
Veuf(veuve)	1,28*	1,06	1,55	1,29*	1,06	1,57	1,30*	1,07	1,58	1,30*	1,07	1,58
Separé(e) / divorcé(e)	1,05	0,77	1,41	1,05	0,76	1,47	1,07	0,75	1,54	1,16	0,79	1,71
Célibataire, jamais marié(e)	0,62*	0,46	0,85	0,62*	0,46	0,85	0,62*	0,46	0,84	0,62*	0,46	0,85
Scolarité												
Pas de diplôme d'études secondaires	1,83*	1,32	2,55	1,83*	1,29	2,61	1,72*	1,19	2,47	2,03*	1,47	2,80
Diplôme d'études secondaires	1,49*	1,08	2,05	1,53*	1,10	2,12	1,41	1,00	1,99	1,63*	1,20	2,21
Études postsecondaires partielles	1,42	0,93	2,15	1,44	0,94	2,19	1,31	0,86	2,01	1,52*	1,02	2,26
Diplôme d'études universitaires†	1,00	...	...	1,00	...	...	1,00	...	...	1,00	...	...
Situation d'emploi												
Personne occupée†	1,00	...	...	1,00	...	...	1,00	...	...	1,00	...	...
Personne en chômage	2,08*	1,63	2,66	2,10*	1,63	2,70	2,10*	1,64	2,70	2,14*	1,65	2,77
Personne inactive	5,32*	4,12	6,86	5,17*	3,94	6,80	5,20*	3,97	6,81	5,29*	4,02	6,96
Quintile de revenu												
Inférieur	1,02	0,74	1,40	1,00	0,72	1,39	0,99	0,72	1,36	1,05	0,76	1,45
Moyen-inférieur	0,93	0,71	1,24	0,92	0,68	1,23	0,90	0,67	1,20	0,95	0,72	1,26
Moyen	0,86	0,64	1,16	0,86	0,64	1,15	0,83	0,62	1,12	0,88	0,66	1,19
Moyen-supérieur	0,98	0,75	1,29	0,98	0,73	1,32	0,97	0,72	1,30	0,99	0,74	1,34
Supérieur†	1,00	...	...	1,00	...	...	1,00	...	...	1,00	...	...
Ascendance autochtone												
Oui	2,85*	1,05	7,69	3,07*	1,11	8,51	2,97*	1,07	8,28	3,49*	1,27	9,58
Non†	1,00	...	...	1,00	...	...	1,00	...	...	1,00	...	...
Minorité visible												
Oui	0,68	0,45	1,03	0,67	0,44	1,02	0,72	0,46	1,13	0,62*	0,41	0,95
Non†	1,00	...	...	1,00	...	...	1,00	...	...	1,00	...	...

† catégorie de référence
* valeur significativement différente de celle observée pour la catégorie de référence ($p < 0,05$)
... n'ayant pas lieu de figurer
Sources : Enquête sur la santé dans les collectivités canadiennes, 2002-2003; Base de données sur la morbidité hospitalière, 2001-2002 à 2004-2005.

lisée et autodéclarée; et 2) évaluer l'effet de l'utilisation d'une situation d'usage modélisée sur des covariables qui sont déjà utilisées pour prédire la situation d'usage du tabac (p. ex. le revenu, la scolarité). On a émis l'hypothèse que la taille de l'effet des covariables pourrait diminuer par suite de l'utilisation de la situation d'usage du tabac modélisée, étant donné que des variables similaires ont également servi à prédire la situation d'usage du tabac.

On a suivi les participants à l'ESCC sur une période de deux ans, à partir de la date de leur entrevue, afin d'identifier celles qui avaient connu au moins une hospitalisation liée à l'usage du tabac, celle-ci se définissant comme une admission liée à un diagnostic principal de maladie respiratoire, de maladie cardiovasculaire ou de cancer (d'après les codes de la CIM-9/10 et de la CIM-10-CA)[25]. On a effectué des analyses de régression logistique pour comparer les résultats de l'utilisation d'une situation d'usage du tabac modélisée plutôt qu'autodéclarée, en fonction des catégories suivantes : personne n'ayant jamais fumé (groupe de référence), personne fumant quotidiennement, et autre. Une approche d'élaboration de modèle a servi à générer des modèles non corrigés (Modèle 1 : situation d'usage du tabac uniquement), des modèles particllement corrigés (Modèle 2 : situation d'usage du tabac plus âge et sexe), et des modèles entièrement corrigés (Modèle 3 : modèle 2 plus variables socioéconomiques).

Les poids d'enquête pour le fichier de l'ESCC couplé ont été rajustés par Statistique Canada, afin de tenir compte de la non-réponse à l'enquête et de l'exclusion des enregistrements des participants à l'enquête qui n'avaient pas consenti au couplage et/ou n'avaient pas fourni de numéro d'assurance-maladie plausible. Pour toutes les analyses, on a tenu compte du plan de sondage complexe de l'enquête en utilisant la méthode du *bootstrap* (rééchantillonnage) pour calculer les estimations de la variance et les intervalles de confiance.

Résultats

Population visée par l'étude

Selon les réponses à l'ESCC de 2000-2001, les personnes n'ayant jamais fumé représentaient environ 41 % de la population à domicile de 25 ans et plus, et les personnes fumant quotidiennement, 26 % de cette population (tableau 1). La majorité des personnes étaient mariées ou vivaient en union libre (71 %), occupaient un emploi (64 %), possédaient leur logement (73 %), vivaient avec au moins une autre personne (85 %), et étaient nées au Canada (76 %). Environ 40 % avaient au moins fait des études postsecondaires partielles. Un peu moins de la moitié (46 %) vivaient dans une région urbaine comptant plus de 500 000 habitants.

Modèles prédictifs

Les variables qui étaient importantes pour prédire la situation d'usage du tabac différaient selon le groupe d'âge et le sexe. Elles sont présentées en ordre d'importance de la signification dans le tableau 2. Dans le cas des modèles prédictifs de la situation de personne fumant quotidiennement, le quintile de revenu, le niveau de scolarité, l'état matrimonial, la propriété du logement et la région du monde du lieu de naissance étaient des prédicteurs significatifs pour tous les groupes âge-sexe. Dans le cas des modèles prédictifs de la situation de personne n'ayant jamais fumé, l'état matrimonial, la propriété du logement, l'ascendance autochtone et la région du monde du lieu de naissance étaient des prédicteurs significatifs pour tous les groupes âge-sexe. Une fois les seuils optimaux propres à l'âge et au sexe appliqués aux probabilités générées pour les modèles prédictifs, près de 80 % des participants à l'enquête ont été affectés à la catégorie des personnes fumant quotidiennement ou à celle des personnes n'ayant jamais fumé, 7,0 % ont été *impossibles à classer*, et 14,6 % ont été classés dans la catégorie *autres*.

Les valeurs de l'ASC allaient de 0,59 à 0,73 et de 0,60 à 0,70, respectivement, dans les modèles de personnes fumant quotidiennement et de personnes n'ayant jamais fumé. Calculé à partir des seuils optimaux, le pourcentage de cas prédits correctement fondé sur des valeurs modélisées allait de 54 % à 67 % pour les personnes fumant quotidiennement, et de 57 % à 65 % pour les personnes n'ayant jamais fumé, les valeurs de l'ASC diminuant avec l'âge. Le pourcentage de cas prédits correctement diminuait lorsqu'on utilisait des bandes plus larges (optimal +/-0,05 et optimal +/-0,10).

Situations d'usage du tabac modélisée et autodéclarée

Des modèles logistiques ont été élaborés pour comparer la contribution des situations d'usage du tabac modélisée et autodéclarée aux fins de la prédiction des hospitalisations liées à l'usage du tabac et de l'évaluation de l'effet de l'utilisation de la variable modélisée à l'égard des covariables qui avaient également servi à prédire la situation d'usage du tabac (p. ex. le revenu, le niveau de scolarité).

Comme il fallait s'y attendre, selon la situation d'usage du tabac *autodéclarée*, le fait d'être une personne fumant quotidiennement plutôt qu'une personne n'ayant jamais fumé était associé à une cote plus élevée exprimant le risque d'au moins une hospitalisation liée à l'usage du tabac, tant dans le modèle non corrigé que corrigé (tableau 3). L'association était similaire, mais plus faible, selon la situation d'usage du tabac *modélisée*. Les rapports de cotes non corrigés pour la situation de personne fumant quotidiennement *modélisée* allaient de 1,81 à 2,99, selon la définition de seuil prise en considération. Les rapports de cotes demeuraient significatifs dans les modèles entièrement corrigés fondés sur le seuil optimal (RC : 1,30) et sur le seuil optimal +/-0,05 (RC : 1,52), mais étaient plus faibles que les rapports de cotes calculés à partir de la situation d'usage du tabac autodéclarée (RC : 2,19).

Dans l'ensemble, les variables comportant une association significative avec les hospitalisations liées à l'usage du tabac dans le modèle reposant sur la situation d'usage du tabac autodéclarée (modèle A) demeuraient significatives dans le modèle fondé sur la situation d'usage du tabac modélisée (tableau 4).

L'âge avancé, l'identité autochtone, le veuvage, le faible niveau de scolarité et le chômage ou le fait de ne pas faire partie de la population active étaient systématiquement associés à une cote plus élevée exprimant le risque d'une hospitalisation liée à l'usage du tabac. Le sexe féminin et le fait de n'avoir jamais été marié(e) étaient associés à une cote plus faible exprimant le risque d'une hospitalisation liée à l'usage du tabac. Le revenu n'était pas associé aux hospitalisations liées à l'usage du tabac, peu importe si le modèle intégrait la situation d'usage du tabac autodéclarée ou modélisée.

Discussion

La présente étude avait pour but d'examiner la faisabilité d'utiliser des techniques de modélisation statistique pour prédire la situation d'usage du tabac, et d'évaluer ensuite l'association entre la variable modélisée et les hospitalisations liées à l'usage du tabac. Les caractéristiques socioéconomiques et démographiques qui étaient prédictives de la situation d'usage du tabac variaient selon l'âge et le sexe, ce qui fait ressortir l'importance d'élaborer des modèles propres à l'âge et au sexe.

L'étape de la validation a révélé des valeurs de l'ASC proches de 0,70 pour la plupart des modèles âge-sexe fondés sur le seuil optimal, celles-ci s'avérant légèrement inférieures aux valeurs obtenues dans d'autres études[26]. Toutefois, le présent projet est unique du fait qu'aucune variable liée à la santé n'a été utilisée pour prédire la situation d'usage du tabac, alors que dans d'autres études, les caractéristiques liées à la santé sont couramment utilisées pour prédire des résultats comme l'hospitalisation et la mortalité. Les valeurs de l'ASC étaient systématiquement faibles pour les modèles des femmes de 65 ans et plus, tant dans le cas des personnes fumant quotidiennement que des personnes n'ayant jamais fumé. La capacité de cerner la situation d'usage du tabac correctement à partir des modèles prédictifs diminuait lorsqu'on élargissait les seuils en en faisant des bandes.

La présente étude est née du besoin d'avoir des données sur les facteurs de risque dans les données tirées du couplage de données de recensement et de données administratives, afin de pouvoir étudier les caractéristiques associées aux résultats pour la santé. Ainsi, il était essentiel de démontrer la faisabilité d'utiliser la situation d'usage du tabac modélisée dans un contexte de recherche, ce que le couplage des données d'enquête et des données hospitalières a permis d'entreprendre.

Les résultats de l'analyse de régression effectuée pour comparer les associations entre la situation d'usage du tabac et les hospitalisations liées à cet usage, pour les situations d'usage modélisée et autodéclarée, démontrent la viabilité d'utiliser la variable modélisée. La situation d'usage du tabac modélisée se comporte de la même façon que la situation d'usage du tabac autodéclarée en ce qui a trait à la direction des associations et à leur signification, quoiqu'elle montre un effet plus faible. Qui plus est, l'utilisation de la situation d'usage du tabac modélisée ne fait pas disparaître les associations entre l'hospitalisation et d'autres covariables (p. ex., l'état matrimonial, la scolarité, la situation d'emploi). L'association entre la situation d'usage du tabac modélisée et l'hospitalisation s'atténue dans les modèles entièrement corrigés, mais demeure significative.

Limites

La présente étude comporte plusieurs limites. L'ESCC exclut des sous-groupes particuliers (membres des Forces canadiennes, habitants des réserves indiennes et de certaines régions éloignées), ainsi que les personnes qui n'ont pas consenti au couplage de leurs données. Ces exclusions peuvent avoir eu un effet sur les modèles finaux utilisés pour prédire la situation d'usage du tabac. La faisabilité d'utiliser la situation d'usage du tabac modélisée a été établie uniquement dans le contexte des hospitalisations liées à l'usage du tabac, au moyen d'une analyse de régression logistique. Un examen plus poussé est nécessaire pour déterminer si cette variable modélisée peut être utilisée dans les études où l'on se sert d'autres techniques (p. ex. l'analyse de la survie) ou résultats (p. ex. la mortalité).

Mot de la fin

Le couplage des données représente une méthode économique pour obtenir des données au niveau de la personne, afin d'étudier les résultats pour la santé au niveau de la population. Toutefois, il se peut que les données comportent des lacunes, plus particulièrement l'absence de données sur les facteurs de risque. La présente étude démontre la faisabilité d'utiliser des techniques de modélisation statistique pour enrichir les sources de données. ■

Références

1. L.L. Roos, J.P. Nicol et S.M. Cageorge, « Using administrative data for longitudinal research: comparisons with primary data collection », *Journal of Chronic Diseases*, 40, 1987, p. 41-49.
2. L.I. Lezzoni, « Using administrative data to study persons with disabilities », *The Milbank Quarterly*, 80(2), 2002, p. 347-379.
3. G.J. Brackstone, « Utilisation des dossiers administratifs à des fins statistiques », *Techniques d'enquête* (n° 12-001 au catalogue), 13, 1987, p. 35-51.
4. J. Disano, J. Goulet, N. Muhajarine *et al.*, « Social-economic status and rates of hospital admission for chronic disease in urban Canada », *The Canadian Nurse*, 106(1), 2010, p. 24-29.
5. G.L. Booth, P. Bisara, L.L. Lipscombe *et al.*, « Universal drug coverage and socioeconomic disparaties in major diabetes outcomes », *Diabetes Care*, 2012, epub ahead of print.
6. R. Wilkins, J.-M. Berthelot et E. Ng, « Tendances de la mortalité selon le revenu du quartier dans les régions urbaines du Canada de 1971 à 1996 », *Rapports sur la santé*, 13(supplément), 2002, p. 51-79.
7. G.E. Hanley et S. Morgan, « On the validity of area-based income measures to proxy household income », *BMC Health Services Research*, 8, 2008, p. 79.
8. C.A. Marra, L.D. Lynd, S.S. Harvard et X. Grubisic, « Agreement between aggregate and individual-level measures of income and education: a comparison across three patient groups », *BMC Health Services Research*, 11, 2011, p. 69.
9. D.A. Southern, L. McLaren, P. Hawe *et al.*, « Individual-level and neighbourhood income measures: agreement and association with outcomes in a cardiac disease cohort », *Medical Care*, 43(11), 2005, p. 1116-1122.
10. R. Pampalon, D. Hamel et P. Gamache, « Une comparaison de données socioéconomiques individuelles et géographiques pour la surveillance des inégalités sociales de santé au Canada », *Rapports sur la santé*, 20(4), 2009, p. 95-105.
11. R.T. Burnett, S. Cakmak, O. Brion *et al.*, « An indirect adjustment method for multiple missing variables applicable to environmental epidemiology », *American Journal of Epidemiology* (en cours d'évaluation).
12. M. Rotermann, « Évaluation de la couverture des données de l'Enquête sur la santé dans les collectivités canadiennes couplées aux dossiers de malades hospitalisés », *Rapports sur la santé*, 20(1), 2009, p. 49-57.
13. K. Wilkins, M. Shields et M. Rotermann, « Utilisation des hôpitaux de soins de courte durée par les fumeurs – Une étude prospective », *Rapports sur la santé*, 20(4), 2009, p. 83-93.
14. C. Sanmartin et S. Khan, *Hospitalisations pour des conditions propices aux soins ambulatoires (CPSA): les facteurs qui importent*, Série de documents de travail de la recherche sur la santé (n° 82-622-X, n° 7, au catalogue), Ottawa, Statistique Canada, 2011.
15. R. Wilkins, M. Tjepkema, C. Mustard et R. Choinière, « Étude canadienne de suivi de la mortalité selon le recensement, 1991 à 2001 inclusivement », *Rapports sur la santé*, 19(3), 2008, p. 27-48.
16. P.A. Peters et M. Tjepkema, « Étude canadienne de suivi de la mortalité et du cancer selon le recensement, 1991 à 2011 », *Symposium 2010, Statistiques sociales : interaction entre recensements, enquêtes et données administratives* (n° 11-522-XCB au catalogue), Ottawa, Statistique Canada, 2011, p. 163-172.
17. Y. Béland, V. Dale, J. Dufour et M. Hamel, « The Canadian Community Health Survey: Building on the success from the past », *Proceedings of the American Statistical Association Joint Statistical Meetings 2005, Section on Survey Research Methods*, Minneapolis, American Statistical Association, août 2005.
18. Statistique Canada, Household Survey Methods Division, *External Linkage of Person-oriented Information 1992/93 to 2000/01 Hospital Morbidity Files* (non publié), Ottawa, Statistique Canada, 2003.
19. C. Nadeau, *Linking HPOI 1992-2005 to CCHS* [documents internes], Division des méthodes d'enquêtes auprès des ménages, Ottawa, Statistique Canada, 2007.
20. S. Wong, M. Shields, S. Leatherdale *et al.*, « Évaluation de la validité de la situation d'usage du tabac autodéclarée », *Rapports sur la santé*, 23(1), 2012, p. 47-54.
21. J.A. Hanley et B.J. McNeil, « The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve », *Diagnostic Radiology*, 143(1), 1982, p. 29-36.
22. C.E. Metz, « Basic principles of ROC analysis », *Seminars in Nuclear Medicine*, 8(4), 1978, p. 283-298.
23. D.L. Streiner et J.C. Cairney, « What's under the ROC? An introduction to Receiver-Operating Characteristics Curves », *The Canadian Journal of Psychiatry*, 52(2), 2007, p. 121-128.
24. J.A. Swets, R.M. Dawes et J. Monahan, « Psychological science can improve diagnostic decision », *Psychological Science in the Public Interest*, 1(1), 2000, p. 1-26.
25. D. Balinuas, J. Patra, J. Rehm *et al.*, « Smoking-attributable morbidity: acute care hospital diagnoses and days of treatment in Canada, 2002 », *BMC Public Health*, 7, 2007, p. 247.
26. L.C. Rosella, D.G. Manuel, C. Burchill et T.A. Stukel, « A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPort) », *Journal of Epidemiology and Community Health*, 65(7), 2011, p. 613-620.